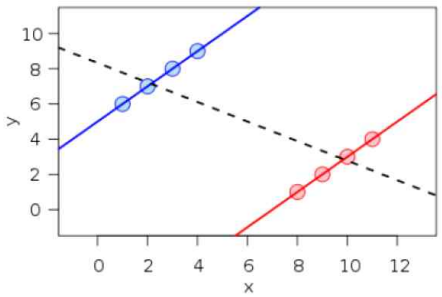
Bio Statistics

1. 인과관계

인과관계와 상관관계는 다르다. 어떤 관계를 분석하기 위해서는 먼저 통계적 관련성을 분석하고, 이것이 인정되면 이것이 인과관계에 의한 것인지 제 3자에 의한 것인지 추가 검정을 해야 알 수 있다. 따라서 단순히 **통계적으로 유의하다는 것이 인과관계라는 뜻은 아니다.**

또한 실험 계획에 의해 수집된 데이터가 아닌 **관찰 데이터에 의한 통계적 분석에는 한계가 있다.**

타이타닉 데이터에서 "승무원들의 생존율은 가장 낮다." 라는 결과가 나왔는데, 이는 성별을 무시한 결과이다. 이는 승무원이 여성에 비하여 더 낮은 생존율을 가진 남자가 대부분 이였기 때문이다. 이처럼 관찰 데이터의 경우 숨겨진 데이터의 영향을 받을 수 있다.



위의 그래프에서 색이라는 범주를 무시하게 되면 X와 Y의 관계는 음의 상관관계처럼 보이지만, 색별로 보면 양의 상관관계임을 알 수 있다.

따라서 통계분석에 의해 인과관계를 설명할 수 있으려면 **숨겨진 요인들의 영향을 제거**하여야 한다.

1. 실험의 목적 : 숨겨진 변수의 영향을 제거하여 인과관계를 판단
2. 실험의 원칙 :
   - **랜덤화** : 실험단위의 배정과 실행 순서를 랜덤화
   - **반복화** : 각 처리를 두 개 이상의 실험단위에 배정
   - **블록화** : 동일한 조건에서 실험하여 실험의 정밀도를 향상
3. 실험단위 : **처리가 이루어지는** 최소 단위
4. 관찰단위 : **관측값이 얻어지는** 최소 단위
5. 측정값 : 참값 + 실험오차 + 관찰오차

▶ 동등한 실험조건과 랜덤화를 통해 숨겨진 변수의 영향을 제거
▶ 반복화를 통해 참값에서 실험오차를 분리

2. 단일모집단 비율

▶ 암과 송전탑과의 관계

$H_0 : p_{\text{삼하리}} = p_0$

$H_1 : p_{\text{삼하리}} \neq p_0, \quad E(\hat{p}) = p_0, \quad Var(\hat{p}) = \frac{p_0(1-p_0)}{n}$

$X \sim B(n,p), \quad \hat{p} = \frac{29}{600} = 0.0041,$

$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1), \quad z = \frac{0.0483 - 0.0041}{\sqrt{\frac{0.0041(0.9959)}{600}}} = 16.956$

$z = 16.956 > z_{0.975} = 1.96$ 이므로 유의수준 5% 하에서 귀무가설을 기각한다. 귀무가설이 기각되므로 시의원의 주장은 통계적으로 타당하다고 볼 수 있다.

구간추정은 때때로 검정보다 더 많은 정보를 제공할 수 있다.

▶ 비율의 신뢰구간(Wald 신뢰구간)

- $\hat{p} + z_{\alpha/2}\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \frac{1}{2}\right)$ (연속성 수정)
- 비율의 구간추정에서 가장 널리 알려진 신뢰구간이다.
- 이해하기 쉽고, 계산도 간단하다는 장점이 있다.
- 그러나, 오류가 많은, 특히 비율의 값이 극단적으로 작을수록 오류가 많은 것으로 알려져 있다.

▶ 비율의 신뢰구간(Wilson 신뢰구간)

- $p_0 \in \left[\frac{\left(\hat{p} + \frac{z^2_{\alpha/2}}{2n}\right) \pm \sqrt{\left(\frac{z^2_{\alpha/2}}{2n} + \hat{p}\right)^2 - \hat{p}^2\left(1 + \frac{z^2_{\alpha/2}}{n}\right)}}{1 + \frac{z^2_{\alpha/2}}{n}}\right]$
- wald 신뢰구간에 비해 p 값이 극단적이여도 오류가 적다.
- 일반적으로 wilson 신뢰구간이 추천된다.

암과 송전탑의 문제의 경우 해당년도의 암환자 발생 비율과 지금까지의 암호나자 발생 비율은 다르다. 단순히 비율 비교를 하는 것이 아니라 **지역에 따른 특성값을 고려**하면 귀무가설을 기각하지 못한다.

3. 단일모집단 평균

Michelson(1882)은 1879년 6월 5일에서 1879년 7월 2일 사이에 포토맥 강변에서 빛이 600m를 왕복하는 시간을 관측하여 빛의 속도를 측정 (빛의 속도 (km/sec)-299,000)

여기서 독립성, 등분산성, 정규성을 가정하고 가설 검정을 실시하고자 한다. 독립성과 등분산성을 확인하기 위해 차분(y_n, y_{n-1})의 그래프를 확인한다. run test와 그래프 모두 귀무가설을 지지한다. 따라서 독립성과 등분산성을 만족한다.

정규성을 확인하기 위해 Shapiro-Wilks 검정을 실시하고, Q-Q Plot을 확인한 결과, 모두 귀무가설을 지지한다. 따라서 정규성을 따른다고 볼 수 있다.

독립성, 등분산성, 정규성을 모두 만족하므로, 평균에 대한 추론을 실시한다. (만약, 정규성을 만족하지 못하는 경우 t-분포는 hotelling's T-분포를 따른다.)

따라서 평균에 대한 일표본 t-test를 실시한 결과 p-value가 2.02e-11이므로 귀무가설을 기각한다. 따라서 Muchelson의 실험은 실제 광속을 측정한 것이라 볼 수 없다.

실험 이후, Newcomb에 의해 3번, Michelson에 의해 1번 더 관측을 실시하였다. 5번의 실험에 대한 분포를 박스플랏을 그려서 보면 특이한 결과를 알 수 있다. **마지막 Michelson의 실험에서는 Newcomb의 실험 결과에 영향을 받은 것**처럼 보여 진다. Newcomb의 실험 역시 점점 진행될수록 **박스플랏의 상자의 크기가 줄어드는 것으로 보아 실험오차가 점점 줄어드는 것**을 알 수 있다.

4. 두 모집단에 대한 추론

절대실험(Absolute Experiment)과 비교실험(Comparative Experiment)

1. 절대실험 : 빛의 속도 측정과 같이 참값이 있으나 측정오차 등으로 측정을 반복해도 동일한 값을 얻지 못하는 상황이다.
2. 상호 비교를 목적으로 하는 실험이다.

두 모집단의 분포에 대한 비교는 가장 흔히 볼 수 있는 통계분석 문제이다. 두 백신의 치료율에 대한 실험1에서 치료율이 각각 0.944(17/18) 과 0.611(11/18) 라고 할 때 첫 번째 백신의 치료율이 높다고 할 수 있는가? A 제약회사에서 생산된 복제약은 오리지날 약과 효능이 같은가?

두 표본에 대한 비율차이 검정의 경우 표본크기가 작을 경우, Wald 신뢰구간에 의한 추론의 경우 위험할 수 있다.

Under H_0

- $\hat{\theta} = (X+Y)/(n+m)$
- $\text{Var}(\hat{\theta}) = \frac{\theta(1-\theta)}{n+m}$

$$z = \frac{\hat{\theta} - 0}{\sqrt{\widehat{\text{Var}}(\hat{\theta})}} \overset{\bullet}{\sim} N(0,1) \quad \text{where } \widehat{\text{Var}}(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n+m}$$

위의 가정 하에서 2 sample prop test 결과 p-value가 0.02251로 귀무가설을 기각한다. 그러나 표본크기가 작아 Wald 신뢰구간에 의한 추론은 오류일 가능성이 높다.

두 표본에 대한 평균 차이 검정의 경우, 등분산성을 만족하지 못할 때

$X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$가 각각 $N(\mu_X, \sigma_X^2)$와 $N(\mu_Y, \sigma_Y^2)$에서 추출된 독립표본

$H_0 : \theta = 0$ vs $H_1 : \theta > 0$ where $\theta = \mu_X - \mu_Y$

$\hat{\theta} = \bar{X} - \bar{Y}$ and $\text{Var}(\hat{\theta}) = \sigma_X^2/m + \sigma_Y^2/n$

$$z = \frac{\hat{\theta} - \theta}{\sqrt{\text{Var}(\hat{\theta})}} \sim N(0,1)$$

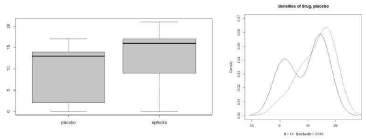
위의 가정 하에서 분산의 추정값은 $\widehat{Var(\hat{\theta})} = \frac{S_X^2}{m} + \frac{S_Y^2}{n}$ 이고, 만약, m과 n이 충분히 크면 근사적으로 자유도가 ν인 t-분포에 근사하는 것으로 알려져 있다.

$$\nu = \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{S_X^4}{m^2(m-1)} + \frac{S_Y^4}{n^2(n-1)}}$$ Welch-Satterthwaite equation

의학실험에서는 표본의 크기가 크지 않는 것이 일반적이다. 따라서 위의 Welch-Satterthwaite에 의한 근사 검정을 사용하기 곤란하다. 평균, 비율, 분산 모두 마찬가지다.

▶ Example

```
placebo = c(0,0,0,2,4,5,13,14,14,14,13,15,17,17)
ephedra = c(0,6,7,9,11,13,16,16,16,17,18,20,21)
boxplot(list(placebo,ephedra),names=c("placebo","ephedra",col="grey")
```



플라시보와 실제 약의 평균 차이 검정을 위해 먼저 각 집단의 가정들을 확인한다. 실제 약(ephedra)의 경우 **약간 오른쪽으로 치우쳐졌으나, 정규성에서는 크게 벗어나지 않지만**, 플라시보(placebo)의 경우 **쌍봉의 형태로 나타났으며 만약 대칭이라면 정규분포라 가정**해도 무리가 없다.

등분산성 역시 var.test(placebo, ephedra) 검정 결과 귀무가설을 기각하지 못하므로 등분산성을 만족한다.

따라서 두 집단에 대한 평균 차이 검정 t.test(placebo, ephedra, var.equal=T)을 실행한 결과 p-value는 0.1291로 귀무가설을 기각하지 못한다. 결론은 **약의 효과가 있는 사람도 존재하나, 평균적으로 약효가 있다는 충분한 근거는 존재하지 않는다.**

만약, t.test(placebo, ephedra, var.equal=F)을 실행한 결과, 즉 **등분산성을 만족하지 못하였을 경우, 해석할 때 Welch의 근사가 만족할 만큼 표본크기가 큰지, 두 모집단에서의 표본크기가 비슷하면, 등분산이 아니여도 등분산 가정하에 실시하여도 큰 문제가 없다. 따라서 실험 설계를 할 때 표본크기가 같도록 설정해야 한다.**

▶ 대응표본 검정
성장환경이 IQ에 미치는 영향을 살펴보기 위해 일란성 쌍둥이를 대상으로 하는 것과 같이 두 집단이 독립적이지 않고 쌍으로 표현될 수 있는 경우에는 대응표본 검정을 실시한다.

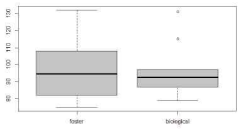
```
> foster = c(80, 88, 75, 113, 95, 82, 97, 94, 132, 108)
> biological = c(90, 91, 79, 97, 97, 82, 87, 94, 131, 115)
> t.test(foster, biological, paired = TRUE)

    Paired t-test

data:  foster and biological
t = 0.041019, df = 9, p-value = 0.9682
alternative: true difference in means is not equal to 0
95 percent confidence interval:
 -5.414902  5.614902
sample estimates:
mean of the differences
                   0.1

> boxplot(list(foster,biological),names=c("foster","biological"))
```

평균에는 차이가 없어 보이나, 상대적으로 foster IQ의 산포가 크다.



실험결과, p-값이 0.9682로 평균의 차이가 없다는 결론이 나왔지만, Boxplot을 보면, 상대적인 분포의 차이를 나타내고 있다.

실험 고려 사항
- 독립 이표본 : 양부모와 생부모에게서 양육된 아이들을 대상으로 각각 표본을 추출하여 IQ Test
- 대응 이표본 : 일란성 쌍둥이 중, 한 아이는 양부모, 한 아이는 생부모에게 양육된 경우에서 표본 추출하여 IQ Test
- 어떤 실험 방법이 맞은 것일까?
- 대응표본은 실험의 결과를 높이기 위해 블록화를 실시한 경우이다.
- 그러나, **블록화**가 항상 가능한 것은 아니다.
- 그러므로, 주어진 조건 하에서 실험의 효과를 최대로 높여야 한다.
- 따라서, **효율적인 통계분석을 위해서는 데이터의 수집단계에서부터 통계적인 요인들을 감안하여 실험을 실시하여야 한다.**

5. 실험계획법 - 실험오차와 관찰오차

빛의 실험을 측정하는 것과 같은 실험을 절대 실험이라고 하고, 두 집단의 차이를 비교하는 실험을 비교실험이라고 한다.

비교실험에서는 관심의 대상인 **요인들의 수준을 몇 가지 설정**하고 각 수준 또는 처리에서 다른 조건들은 동일하게 설정하여 실험을 한 후, 요인 수준에 따른 반응값들이 통계적으로 유의한가를 검증한다.

▶ Example : 금붕어 사료에 따른 몸무게의 증가분
1. **실험방법_1** : 금붕어 10마리가 들어 있는 어항 3개에 각각 다른 종류의 사료를 주어 일정 기간 후에 몸무게를 측정한다.
   - 실험단위 : 어항
   - 관찰단위 : 금붕어

2. **실험방법_2** : 금붕어 1마리가 들어 있는 어항 30개에 각각 10개 씩 다른 종류의 사료를 주어 일정 기간 후에 몸무게를 측정
   - 실험단위 : 어항(금붕어)
   - 관찰단위 : 금붕어(어항)

3. 실험단위와 관찰단위에서 각각 오차(error)가 발생한다.

실험오차 : 동량의 사료를 준다고 해도 실제 사료량은 다를 수 있음.
어항의 위치에 따라 일조량, 소음 등의 다른 요인이 개입함.
관찰오차 : 측정오차
금붕어에 따라 성장률이 다를 수 있음.

측정값 = 참값 + 실험오차 + 관찰오차

4. 측정값에서 오차와 참값을 분리하기 위해서는 실험의 원칙이 준수되어야한다.

▶ 실험의 원칙
1. 랜덤화 : 실험의 객관성을 보장하기 위해 실험 단위의 배정 및 실험 순서를 랜덤하게 결정, 만약 **랜덤화를 하지 않을 경우 숨겨진 변수의 영향이 작동할 가능성이 높다.**

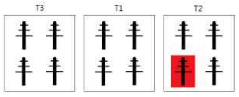
2. 반복화 : 각각의 처리를 둘 이상의 실험단위에 대해 실험, 반복화를 통해 **실험오차가 서로 average out 되기를 기대한다.**

3. 블록화 : 실험의 성격상 들어난 요인이 있을 경우 그 요인이 동일하도록 실험 조건을 설정, 블록화는 실험의 원칙이라기 보다는 **실험의 정밀도를 향상시키기 위한 기법**이다.

▶ 실험계획법 예제
소나무 묘목의 성장에 영향을 미치는 공해물질의 효과를 연구하고자 한다.

요인은 level 1(T1) : 정화된 공기 / level 2(T2) : 이산화황 / level 3(T3) : 이산화질소 로 구분한다.

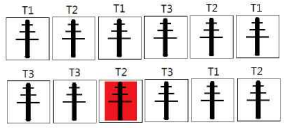
12구루의 묘목을 마련하여 3개의 온실에 각각 4개씩 랜덤하게 배치하고, 각 온실마다 처리를 랜덤하게 배치한다.



$$y_{ij} = \mu + \tau_i + \epsilon_i + \eta_{ij}$$

μ = 전체 평균, τ_i = I번째 처리 효과, ϵ_i = 실험오차, η_{ij} = 관측오차

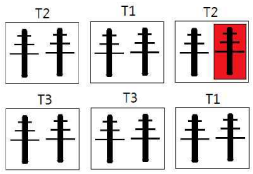
빨간색으로 표시된 값은 2번째 처리의 3번째 관측값이므로 y_{23} 이다. 위의 실험 모델은 **실험오차와 측정오차가 교락되어 분리할 수 없다.** 따라서 묘목의 성장의 차이가 정확하게 처리 때문인지, 실험오차인지는 알 수 없다. **반복화가 준수되지 않은 실험 설계이다.**



$$y_{ij} = \mu + \tau_i + \epsilon_i + \eta_{ij} = \mu + \tau_i + e_{ij}$$

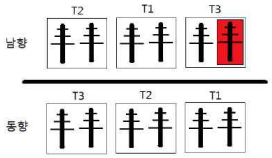
빨간색으로 표시된 값은 2번째 처리의 3번째 관측값이므로 y_{23} 이다. 이번 실험은 **실험오차와 관찰오차가 교락 되었으나 처리효과에만 관심이 있으므로 분리할 필요가 없다.**

위와 같은 모형을 **완전임의배치법(CRD : Completely Randomized Design)**이라고 부른다.



$$y_{ij} = \mu + \tau_i + \epsilon_i + \eta_{ijk} = \mu + \tau_i + e_{ijk}$$

빨간색으로 표시된 값은 2번째 처리의 2번째 온실의 2번째 값이므로 y_{222} 로 표현한다. 처리효과, 실험오차, 관찰오차가 모두 분리되었지만, 두 번째 설계가 일반적으로 검증력이 더 높다.



$$y_{ij} = \mu + \tau_i + \epsilon_i + \eta_{ijk} = \mu + \tau_i + e_{ijk}$$

성장률이 일조량과 밀접한 관계가 있다는 사실을 실험 전에 알고 있을 때 더 정확한 실험을 위해 일조량이 같은 온실에서 실시해야 한다. 따라서 **남향 3개와 동향 3개에 각각 3개의 처리를 랜덤하게 배치한다.(블럭화)**

빨간색으로 표시된 값은 3번째 처리의 1번째 블럭의 2번째 값이므로 y_{312} 로 표현한다. 위의 실험은 처리효과와 실험오차, 관찰오차가 모두 분리되었다. 이런 모형을 **랜덤블럭화모형(RBD : Random Block Design)**이라 한다.

6. 실험계획법 – 완전임의배치법

▶ 1요인 완전임의배치법(one-way CRD)
4개의 브랜드의 타이어 마모도 차이의 대한 실험이다. 동일 종류의 차에 타이어를 부착하고 20,000km 주행 후, 타이어의 지름을 측정하여 마모된 정도를 측정한다.

car			
I	II	III	IV
A	B	C	D
A	B	C	D
A	B	C	D
A	B	C	D

우리는 일반적으로 실험을 이렇게 계획할 것이다. 그러나 이 방법에는 여러 가지 문제가 있다. 각 자동차들마다 마모 정도의 차이를 인지하지 못한 실험 계획이다. (실험오차와 처리효과가 교락된 상태다.)

따라서 16개의 타이어를 랜덤하게 16개의 위치에 부착하는 CRD 모형을 설계한다.

car			
I	II	III	IV
C(12)	A(14)	D(10)	A(13)
A(17)	A(13)	C(11)	D(9)
D(13)	B(14)	B(14)	B(8)
D(11)	C(12)	B(13)	C(9)

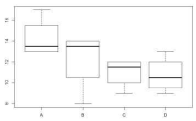
Brand				
A	B	C	D	
14	14	12	10	
13	14	11	9	
17	8	12	13	
13	13	9	11	
$\bar{y}_{\cdot j}$ 14.25	12.5	11	10.75	$\bar{y}_{\cdot\cdot} = 12.125$

- 모형(일원배치 분산분석모형)

$$y_{ij} = \mu_j + \epsilon_{ij}, i = 1,2,3,4; j = 1,2,3,4$$
$$\text{또는 } y_{ij} = \mu + \tau_j + \epsilon_{ij}, i = 1,2,3,4; j = 1,2,3,4$$
$$\text{where } \epsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2) \text{ and } \tau_j = \mu_j - \mu$$

그러나 이 실험 역시 μ 와 τ_j 는 추정이 불가능하다. 각 자동차별의 실험오차와 타이어의 처리효과를 완벽하게 분리하여 추정할 수 없기 때문이다. 그러나 검정 방법은 동일하다.

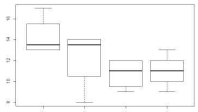
```
abrasion <- c(14,14,12,10,13,14,11,9,17,8,12,13,
              13,13,9,11)
brand <- as.factor(rep(c("A", "B", "c", "D"),4))
tire <- data.frame(abrasion, brand)
abrasion.aov <- aov(abrasion ~ brand, data = tire)
summary(abrasion.aov)
abrasion.glm <- glm(abrasion ~ brand, data = tire)
summary(abrasion.glm)
boxplot(abrasion ~ brand, data=tire)
```



```
            Df Sum Sq Mean Sq F value Pr(>F)
brand        3  30.69  10.229   2.443  0.115
Residuals   12  50.25   4.188
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.250      1.023  13.927 9.05e-09 ***
brandB        -2.000      1.447  -1.382   0.1921
brandC        -3.250      1.447  -2.246   0.0443 *
brandD        -3.500      1.447  -2.419   0.0324 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
abrasion = c(17,14,12,13,14,14,12,11,13,13,10,
             11,13,8,9,9)
brand=as.factor(rep(c("A", "B", "C", "D"),4))
car = as.factor(rep(1:4, each=4))
tire = data.frame(abrasion, brand, car)
boxplot(abrasion ~ brand, data=tire)
abrasion.aov=aov(abrasion~car+brand,data=tire)
summary(abrasion.aov)
abrasion.glm=glm(abrasion~car+brand,data=tire)
summary(abrasion.glm)
```



```
            Df Sum Sq Mean Sq F value  Pr(>F)
car          3  38.69  12.896  10.038 0.00313 **
brand        3  30.69  10.229   7.962 0.00668 **
Residuals    9  11.56   1.285
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.1875     0.7497  21.592 4.62e-09 ***
car2         -1.2500     0.8015  -1.560 0.153279
car3         -2.2500     0.8015  -2.807 0.020466 *
car4         -4.2500     0.8015  -5.303 0.000492 ***
brandB       -2.0000     0.8015  -2.495 0.034118 *
brandC       -3.5000     0.8015  -4.367 0.001805 **
brandD       -3.2500     0.8015  -4.055 0.002863 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

▶ 고정요인과 랜덤요인
이번 실험에서는 관심 대상인 브랜드가 4개이다. 그러므로 브랜드 효과는 고정효과이다. ANOVA 결과 브랜드의 처리 효과가 유효하다면 4개의 브랜드 사이에는 효과가 있다고 말할 수 있다. 그러나 다른 브랜드에 대해서는 모른다. 만약 브랜드가 랜덤요인이라면 브랜드가 효과가 있다는 것을 의미하므로 실험되지 않는 4개의 브랜드 외에도 확대 해석이 가능하다.

그러나 어떤 요인이 랜덤인지 고정인지 판단하는 것은 쉽지 않다. 자동차의 경우 만약 수많은 차 중에서 4개를 랜덤을 뽑았다면 랜덤효과이고, 가용할 수 있는 차가 4개 밖에 없는 상태였다면 고정효과이다.

블록효과는 개념상 랜덤효과지만 검정의 대상이 아니다.

랜덤요인에 대해서는 확대 해석이 가능하므로 기각을 하기 위해서는 보다 강한 증거가 필요하다. 즉 검정 방법이 다르므로 모형 추정에서 랜덤효과를 지정해 주어야 한다.

▶ 라틴 방격 모형

앞선 RBD 모형에서 자동차를 블록화함으로써 차의 영향을 average out 하기를 기대하는 모형을 설계했다. 여기서 타이어의 부착 위치도 마모에 영향을 미치는 요인이라고 생각되어 블록화를 할 수 있다.

즉 각 Brand가 모든 차에 하나씩 장착되어야 하고, 각기 다른 4개의 위치에 하나씩 장착되어야 한다. 이러한 모형을 라틴방격모형이라고 한다.

	car				
Position	I	II	III	IV	$y_{\cdot\cdot k}$
1	C(12)	D(11)	A(13)	B(9)	44
2	B(14)	C(12)	D(11)	A(13)	50
3	A(17)	B(14)	C(10)	D(9)	50
4	D(13)	A(14)	B(13)	C(9)	49
$y_{i\cdot\cdot}$	56	51	47	39	193

$$y_{ijk} = \mu + \beta_i + \tau_j + \gamma_k + \epsilon_{ijk}$$

Do not cover this topic.

7. 실험계획법 – 랜덤화 블록 설계

위의 CRD 모형에서의 문제점은 다음과 같다.

1. A 타이어는 3번째 차에서, B 타이어는 첫 번째 차에서는 실험되지 않았다.
2. 오차항은 실험오차뿐만 아니라 차의 효과에 의한 변동을 포함한다.
3. 오차항에서 차의 효과에 의한 변동을 제거하여 실험오차를 줄여야 한다.
4. RBD 모형을 통해 차의 효과가 averaged out 되기를 기대하지만, 완벽하게 제거되는 것은 아니다. 따라서 각 브랜드를 모든 차에서 실험하게 계획한다.

car			
I	II	III	IV
B(14)	D(11)	A(13)	C(9)
C(12)	C(12)	B(13)	D(9)
A(17)	B(14)	D(11)	B(8)
D(13)	A(14)	C(10)	A(13)

Brand					
Car	A	B	C	D	$y_{i\cdot}$
I	17	14	12	13	14
II	14	14	12	11	12.75
III	13	13	10	11	11.75
IV	13	8	9	9	9.75
$\bar{y}_{\cdot j}$	14.25	12.5	10.75	11	$\bar{y}_{\cdot\cdot} = 12.0625$

- 모형 (2원배치, but 1-factor)
- $y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}, i = 1,2,3,4; j = 1,2,3,4, \quad \epsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$
  β_i: 차(block) 효과; τ_j: 처리(brand) 효과
- As before we want to test

$$H_0 : \tau_j = 0 \text{ for all } j \text{ vs } H_0 : \tau_j \neq 0 \text{ for some } j$$

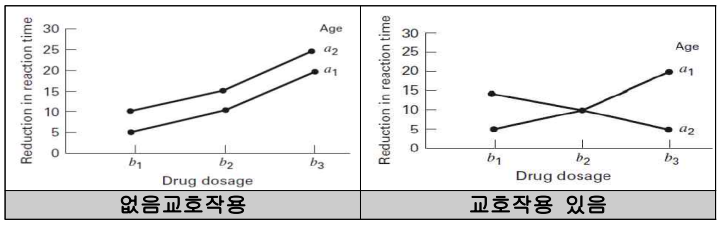
8. 실험계획법 – 2 요인 실험

요인 A : 투여량, 10mg, 15mg, 20mg (level 3)
요인 B : 연령, 65세 미만, 65세 이상 (level 2)

실험 결과에 영향을 미치는 요인이 2개인 경우, 예를 들어 어떤 약의 투여량과 연령에 따라 반응 시간이 다르지를 실험하는 경우 2요인의 결합으로 6개의 수준이 나타난다.

$$y_{mk} = \mu + \tau_m + \epsilon_{mk}, \epsilon_{mk} \overset{iid}{\sim} N(0,\sigma_\epsilon^2)$$
$$m = 1,\ldots,6; k = 1,\ldots,n \text{(각 처리마다 실험횟수가 동일하다고 가정)}$$
$$\text{첨자 } m \text{을 두 요인의 수준에 따라 } (i,j) \text{로 분리}$$
$$y_{ijk} = \mu + \tau_{ij} + \epsilon_{ijk}, \epsilon_{ijk} \overset{iid}{\sim} N(0,\sigma_\epsilon^2)$$
$$i = 1,2,3; j = 1,2; k = 1,\ldots,n$$

▶ 교호작용 확인
A의 각 수준에서 B의 효과가 동일한지 확인을 해야 한다.



만약 오른쪽의 모형처럼 **교호작용이 나타났다면 모형에 교호작용을 포함시키고 분석**해야 한다.

▶ 분지 모형 (Nested Design)
한 요인 안에 다른 요인이 들어가 있어 분지 모형처럼 보이는 것이 특징이다.

전국의 상수원마다 불소함유량이 같은지를 조사

- 상수원 3개를 임의 선택, 각 상수원에서 3 곳씩 취수지를 임의 선정하고, 표본 (1 리터)을 두번 채취하여 각각 불소함유량 측정

상수원	1			2			3		
취수지	1	2	3	1	2	3	1	2	3
함유량	y_{111} y_{112}	y_{121} y_{122}	y_{131} y_{132}	y_{211} y_{212}	y_{221} y_{222}	y_{231} y_{232}	y_{311} y_{312}	y_{321} y_{321}	y_{331} y_{332}

$$y_{ijk},\ i(\text{상수원}) = 1,2,3;\ j(\text{취수지}) = 1,2,3;\ k(\text{표본}) = 1,2$$

측정값에 영향을 미치는 요인

- A: 상수원, B: 취수지

비록 1번 상수원의 1번 취수지와 2번 상수원의 1번 취수지는 다른 장소이다. 따라서 y_{111} 과 y_{211} 는 두 번째 첨자가 같지만, 다른 장소이다. **분지 모형은 B가 A에 지분된 것이기 때문에 교호작용이 무의미**하다.

$$y_{ijk} = \mu + A_i + B_{j(i)} + \epsilon_{k(ij)}$$

▶ 4요인 실험계획법 (분지가 포함)
- A,B,C,D 4개의 요인이 존재
- B와 C는 A에서 각각 지분되었으며, 서로 교차이다.

```
gunData <- c(20.2, 26.2, 23.8, 22.0, 22.6, 22.9, 23.1, 22.9, 21.8,
24.1, 26.9, 24.9, 23.5, 24.6, 25.0, 22.9, 23.7, 23.5,
14.2, 18.0, 12.5, 14.1, 14.0, 13.7, 14.1, 12.2, 12.7,
16.2, 19.1, 15.4, 16.1, 18.1, 16.0, 16.1, 13.8, 15.1)
method <- factor(rep(1:2, each = 18))
G <- factor(rep(rep(1:3, each = 3), times = 4))
team <- factor(rep(1:3, times = 12))
canonloading <- data.frame(gunData, method, group, team)
str(canonloading)

'data.frame':   36 obs. of  4 variables:
 $ gunData: num  20.2 26.2 23.8 22 22.6 22.9 23.1 22.9 21.8 24.1 ...
 $ method : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ G      : Factor w/ 3 levels "1","2","3": 1 1 1 2 2 2 3 3 3 1 ...
 $ team   : Factor w/ 3 levels "1","2","3": 1 2 3 1 2 3 1 2 3 1 ...

library(EMSaov)
result <- EMSanova(gunData ~ method + G + team, type = c("F", "F", "R"),
                   nested = c(NA, NA, "G"), data = canonloading)
result
             Df     SS     MS     F     P Sig                                EMS
method        1 651.95 651.95 364.8 <0.00 *** E+2method:team(G)+18method
G             2  16.05   8.02   1.2  0.35              E+4team(group)+12G
method:G      2   1.18   0.59   0.3  0.72 E+2method:team(G)+6method:G
team(G)       6  39.25   6.54   2.8  0.04   *                  E+4team(G)
method:team(G)  6  10.72   1.78   0.7  0.60            E+2method:team(G)
Residuals      18  41.59   2.31                                       E
```

모형설정을 위한 3 step

1. 모든 요인을 콜론과 지분된 요인과 같이 표시
   - A:, B:A, C:A, D:
2. 모든 가능한 symbolic product를 구하되, 콜론 오른쪽에 중복 기호는 하나의 기호로 표시, 콜론의 양쪽에 같은 기호가 있는 항은 제거
   - AB:A, AC:A, AD:, BC:A, BD:A, CD:A, ABC:A, ABD:A,BCD:A
   - AD: BC:A BD:A BCD:A
3. 1.의 주효과와 2.의 교호작용을 모형으로 표시: 요인별 첨자, i, j, k, l, m

$$y_{ijklm} = \mu + A_i + B_{j(i)} + C_{k(i)} + D_l + AD_{il} + BC_{jk(i)} + BD_{jl(i)} + BCD_{jkl(i)} + \epsilon_{m(ijkl)}$$

$$y_{ijkm} = \mu + A_i + B_{j(i)} + C_k + BC_{jk(i)} + \epsilon_{m(ijk)}$$

각 요인의 수준, n_a, n_b, n_c, n_w

- 분산분석표 작성을 위한 규칙
  - Rule 1: 모형에 나타난 모든 요인을 요인별로 한 줄에 표시
  - Rule 2: $AB : XY$의 자유도는 $(n_a - 1)(n_b - 1)n_x n_y$, 총 제곱합의 자유도는 $n_a n_b n_c m_w - 1$
  - Rule 3: $AB : XY$의 SS는 $(a-1)(b-1)xy$, 총제곱합은 $abcw - 1$

ANOVA table

Source	df	ss	MS	F	EMS
A	$n_a - 1$	$a - 1$			
$B : A$	$(n_b - 1)n_a$	$(b-1)a$			
C_k	$n_c - 1$	$(c-1)$			
$BC : A$	$(n_b - 1)(n_c - 1)n_a$	$(b-1)(c-1)a$			
$W : ABC$	$(n_w - 1)n_a n_b n_c$	$(w-1)abc$			
Total	$n_a n_b n_c n_w - 1$	$abcw - 1$			

  - Rule 4: $B : A$의 자유도 $= (n_b - 1)n_a = n_a n_b - n_a$
  - Rule 5: SS에서 $1 = N\bar{y}^2_{\cdots}, (b-1)a = ab - a$ where $N = n_a n_b n_c n_w$ and

$$ab = \frac{\sum_{i=1}^{n_a}\sum_{j=1}^{n_b}\left(\sum_{k=1}^{n_c}\sum_{m=1}^{n_w} y^2_{ijkm}\right)}{n_c n_w}$$

- 랜덤 또는 고정 요인인지에 관계없이 자유도와 SS는 동일
- 각 SS를 자유도로 나누어 MS를 계산
- MS의 기대값인 EMS는 요인의 종류에 따라 다르고, 검정을 위한 F 값도 다르게 계산된다.
- 즉, F 값은 EMS를 참조하여 계산된다.

▶ 다요인 계획법 예제

해전에서 대표의 발사 속도는 승패를 결정하는 중요한 요인인 바, 대포의 발사 속도를 높히기 위해 포탄을 대포에 loading하는 방법에 대해 연구

- 새로운 loading 방법을 기존의 방법과 비교 실험, 요인 M: loading 방법 (고정),
- loading하는 포병의 체격조건에 따라 세 개의 그룹(slight, average, heavy)으로 나누어 실험, 요인 G: 체격 조건 (고정)
- 각 그룹 별로 3 개의 팀, 총 9 개의 팀을 임의 선정, 요인 T: 팀(랜덤)
- 팀 별로 각 loading방법을 2 번씩 실험

Group	I			II			III		
Team	1	2	3	4	5	6	7	8	9
Method I	20.2 24.1	26.2 26.9	23.8 24.9	22.0 23.5	22.6 24.6	22.9 25.0	23.1 22.9	22.9 23.7	21.8 23.5
Method II	14.2 16.2	18.0 19.1	12.5 15.4	14.1 16.1	14.0 18.1	13.7 16.0	14.1 16.1	12.2 13.8	12.7 15.1

방법, 그룹, 팀의 처리효과와 방법과 그룹, 방법과 팀의 교호작용이 존재한다. (그룹과 팀은 분지이므로 교호작용이 의미가 없다.)

- $M:, G:, T : G, MG:, MT : G$
- 모형

$$y_{ijkm} = \mu + M_i + G_j + MG_{ij} + T_{k(j)} + MT_{ik(j)} + \epsilon_{m(ijk)}$$

$$T_{k(j)} \overset{iid}{\sim} N(0, \sigma^2_T),$$

$$i = 1,2(n_M); j = 1,2,3(n_G); k = 1,2,3(n_T); m = 1,2(n_w)$$

Source	df	SS	MS	F	EMS
M	1	M-1			$\sigma^2_\epsilon + 2\sigma^2_{MT} + 18\phi_M$
G	2	G-1			$\sigma^2_\epsilon + 4\sigma^2_T + 12\phi_G$
MG	2	(M-1)(G-1)			$\sigma^2_\epsilon + 2\sigma^2_{MT(G)} + 6\phi_{MG}$
$T(G)$	6	(T-1)G			$\sigma^2_\epsilon + 4\sigma^2_T$
$MT(G)$	6	(M-1)(T-1)G			$\sigma^2_\epsilon + 2\sigma^2_{MT(G)}$
오차	18	(W-1)MGT			σ^2_ϵ

- Test for $H_0 : M_i = 0$ for all i vs. $H_1 : M_i \neq 0$ for some i
- $F = MS_M / MS_{MT:G}$ on 1, 6 df

R로 실행 ANOVA는 다음과 같다.

```
gunData <- c(20.2, 26.2, 23.8, 22.0, 22.6, 22.9, 23.1, 22.9, 21.8,
24.1, 26.9, 24.9, 23.5, 24.6, 25.0, 22.9, 23.7, 23.5,
14.2, 18.0, 12.5, 14.1, 14.0, 13.7, 14.1, 12.2, 12.7,
16.2, 19.1, 15.4, 16.1, 18.1, 16.0, 16.1, 13.8, 15.1)
method <- factor(rep(1:2, each = 18))
G <- factor(rep(rep(1:3, each = 3), times = 4))
team <- factor(rep(1:3, times = 12))
canonloading <- data.frame(gunData, method, group, team)
str(canonloading)

'data.frame':   36 obs. of  4 variables:
 $ gunData: num  20.2 26.2 23.8 22 22.6 22.9 23.1 22.9 21.8 24.1 ...
 $ method : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ G      : Factor w/ 3 levels "1","2","3": 1 1 1 2 2 2 3 3 3 1 ...
 $ team   : Factor w/ 3 levels "1","2","3": 1 2 3 1 2 3 1 2 3 1 ...

library(EMSaov)
result <- EMSanova(gunData ~ method + G + team, type = c("F", "F", "R"),
                   nested = c(NA, NA, "G"), data = canonloading)
result
             Df     SS     MS     F     P Sig                                EMS
method        1 651.95 651.95 364.8 <0.00 *** E+2method:team(G)+18method
G             2  16.05   8.02   1.2  0.35              E+4team(group)+12G
method:G      2   1.18   0.59   0.3  0.72 E+2method:team(G)+6method:G
team(G)       6  39.25   6.54   2.8  0.04   *                  E+4team(G)
method:team(G)  6  10.72   1.78   0.7  0.60            E+2method:team(G)
Residuals      18  41.59   2.31                                       E
```

Source	EMS
M_i	$\sigma^2_\epsilon + 18\phi_M$
G_j	$\sigma^2_\epsilon + 4\sigma^2_T + 12\phi_G$
$T_{k(j)}$	$\sigma^2_\epsilon + 4\sigma^2_T$
MG_{ij}	σ^2_ϵ
$MT_{ik(j)}$	σ^2_ϵ
$\epsilon_{m(ijk)}$	σ^2_ϵ

  - The EMS's of Method * Group, Method * Team and Error are identical.
  - Combine the SS's of Method * Group and Method * Team with Error
- Our final model would be

$$y_{ijkm} = \mu + M_i + G_j + T_{k(j)} + \epsilon_{m(ijk)}$$

with $i = 1, 2$; $j = 1, 2, 3$; $k = 1, 2, 3$ for all j and $m = 1, 2$ for all i, j, k.

검정결과 방법과 그룹, 방법과 팀에 대한 교호작용은 의미가 없는 것으로 나타났다.